Transportation Recommendation with Fairness Consideration

Ding Zhou[1], Hao Liu[2(✉)], Tong Xu[1], Le Zhang[1], Rui Zha[1], and Hui Xiong[3(✉)]

[1] School of Computer Science and Technology, University of Science and Technology of China, Hefei, China
{zhouding,laughing,zr990210}@mail.ustc.edu.cn, tongxu@ustc.edu.cn

[2] Business Intelligence Lab, Baidu Research, Beijing, China
liuhao30@baidu.com

[3] Rutgers University, New Brunswick, USA
hxiong@rutgers.edu

Abstract. Recent years have witnessed the widespread use of online map services to recommend transportation routes involving multiple transport modes, such as bus, subway, and taxi. However, existing transportation recommendation services mainly focus on improving the overall user click-through rate that is dominated by mainstream user groups, and thus may result in unsatisfactory recommendations for users with diversified travel needs. In other words, different users may receive unequal services. To this end, in this paper, we first identify two types of unfairness in transportation recommendation, (*i*) the *under-estimate* unfairness which reflects lower recommendation accuracy (*i.e.*, the quality), and (*ii*) the *under-recommend* unfairness which indicates lower recommendation volume (*i.e.*, the quantity) for users who travel in certain regions and during certain time periods. Then, we propose the **F**airness-**A**ware **S**patiotemporal **T**ransportation **R**ecommendation (**FASTR**) framework to mitigate the transportation recommendation bias. In particular, based on a multi-task wide and deep learning model, we propose the dual-focal mechanism for under-estimate mitigation and tailor-designed spatiotemporal fairness metrics and regularizers for under-recommend mitigation. Finally, extensive experiments on two real-world datasets verify the effectiveness of our approach to handle these two types of unfairness.

Keywords: Transportation recommendation · Personalized recommendation · Fairness machine learning

1 Introduction

Transportation recommendation is a one-stop routing service, which aims to help users find the most proper transport mode (*e.g.*, bus, subway, and taxi) and combinations, by given the Origin-Destination pair of users. As an emerging map service in various online navigation applications (*e.g.*, Baidu Maps, Google Maps), transportation recommendation has deeply penetrated the citizens' daily

C. S. Jensen et al. (Eds.): DASFAA 2021, LNCS 12683, pp. 566–578, 2021.
https://doi.org/10.1007/978-3-030-73200-4_40

lives. For instance, the transportation recommendation service on Baidu Maps is answering over ten million queries made by millions of users in China per day.

Due to the practicality of transportation recommendation, there has been an increasing attention to this field from both academia and industry. Recently, different strategies are proposed to recommend transport modes for users, such as historical trajectories based strategy [20], shortest distance based strategy [10] and city graph based strategy [13,14,19]. Although existing works can achieve good performance in transportation recommendation, they overlook two types of unfairness that we observe from large-scale historical recommendation log. One is under-estimate unfairness, which may lead to lower recommendation accuracy on minorities. Since the majority loss functions minimize the overall error of model that benefits mainstream groups, this under-estimate unfairness (*e.g.*, big performance gap between different transport modes) is becoming increasingly prevalent. The other is the under-recommend unfairness, which may lead to lower recommendation volume for minorities' transportation needs. For instance, bus and subway that concentrated in the center of the city are the protagonists during rush hour, which may greatly squeeze the recommendation volume of other transport modes like taxi. In other words, users who live in suburban and need taxi at that time can not be recommended and satisfied. Furthermore, these two types of unfairness may increase homogeneity and decrease utility [5] of the recommender services.

Recently, the machine learning fairness community primarily focuses on fairness in classification and has proposed various definitions of fairness [3,17], such as group fairness [4,8] that restricts any two groups to having equal probability of being assigned to the positive predicted class, and equality of opportunity [12] that restricts any two groups to having equal false negative rate. For an unbiased recommendation, [1] and [2] focus on fairness in pointwise and pairwise accuracy of learning to rank, respectively. However, these fairness metrics can not satisfy the spatiotemporal settings in transportation recommendation. Therefore, a more comprehensive solution is still urgently required for these challenges.

To that end, we propose the Fairness-Aware Spatiotemporal Transportation Recommendation (FASTR) framework for effective and fair transportation recommendation. Specifically, we first introduce a wide and deep learning model [7] modified with multi-task mechanism for capturing feature co-occurrence and high-order interaction relationships. Besides, we propose a dual-focal mechanism to mitigate under-estimate unfairness, which consists of task-level focal loss for enhancing the prediction of each individual task and relation-level focal loss for mitigating performance gap between tasks. Then we propose multiple well-designed spatiotemporal fairness metrics to quantify the under-recommend unfairness in certain regions and time periods. Furthermore, with the help of the proposed fairness metrics, a series of tailor-designed regularizers are proposed to guide the optimization for mitigating the under-recommend unfairness.

Overall, the major contributions of our work can be summarized: 1) To the best of our knowledge, our FASTR model is among the first product-level intelligent transportation recommender that focuses on mitigating under-estimate and under-recommend unfairness, 2) We utilize multi-task wide and deep model with the well-designed dual-focal loss for under-estimate unfairness mitigation, and we propose tailor-designed spatiotemporal fairness metrics and regularizers to mitigate under-recommend unfairness, 3) Extensive experiments on real-world datasets verify the effectiveness of our approach on handling under-estimate and under-recommend unfairness.

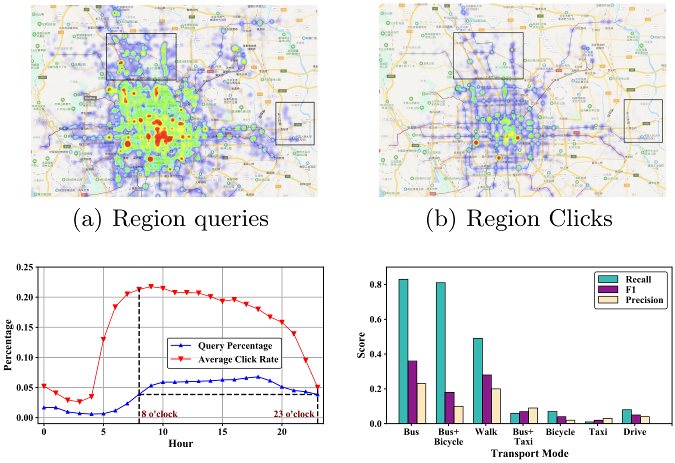
2 Data Description and Analysis

In this section, we first introduce the datasets and the constructed features used in our work, and we analyze how unfairness appears in transportation recommendation subsequently. Specifically, we collected our datasets from Baidu Maps, a large-scale navigation application, from July 2019 to September 2019 in Beijing and Shanghai. And according to user interaction loop, our source data $\mathcal{D}$ can be further categorized into query records, click records and the corresponding context features. In short, for each sample in our datasets $\mathcal{D}$, its query record represents one transportation search (*e.g.*, Origin-Destination pair) from a user on Baidu Maps, and its click record indicates the user's feedback on different recommendations (*e.g.*, a user may click on specific transportation recommendation for him/her). Meanwhile, the corresponding context features for each sample consist of spatial features, temporal features, meteorological features, user features and transport mode features, where the details are shown in Table 1. Totally, we have 5,327,897 samples with 1,177,844 clicks in Beijing, as well as 5,120,561 samples with 1,190,813 clicks in Shanghai. And for each sample, we consider 7 transport modes can be recommended in datasets $\mathcal{D}$ (*Bus*, *Bus + Bicycle*, *Walk*, *Bus + Taxi*, *Bicycle*, *Taxi* and *Drive*).

Table 1. Corresponding context features for each sample

Feature	Composition
Spatial features	*District category*, *Point-Of-Interest (POI) category*, *POI count*, *Transport Mode Click count*
Temporal features	*Hour*, *Minute*, *Day of Week*, *Day of Month*, *Workday*
Meteorological features	*Weather*, *Temperature*, *Air Quality Index*, *Wind Speed*, *Wind Direction*
User features	*Demographic Attribute*, *Social Attribute*, *User Historical Transport Mode Distribution*
Transport mode features	*Price*, *Time*, *Distance*

To further understand the unfairness phenomenon in transportation recommendation, we analyze our datasets from under-estimate and under-recommend aspects in Beijing with our original model [15]. Note that we have similar observations in Shanghai. Since user clicks can proxy the recommendation accuracy



(a) Region queries

(b) Region Clicks

(c) Temporal Clicks

(d) Different Performance

Fig. 1. Distribution and performance of Beijing dataset. (a) queries distribution in region aspect; (b) clicks distribution in region aspect; (c) temporal distribution of average query percentage and clicks rate per hour; (d) the precision, f1-score and precision performance on different transport modes in Beijing.

and volume, we first calculate the distribution of click rate in different regions and time periods to reveal the unfairness. As shown in Fig. 1(a) and Fig. 1(b) that depict the region distribution of queries and clicks respectively in Beijing, the click rate in rectangles of Fig. 1(a) and Fig. 1(b) is much lower or even close to zero compared with other regions, which indicates under-estimate and under-recommend unfairness happened in certain regions. In Fig. 1(c), we can see that the average click rate at 23 o'clock is much lower than 8 o'clock even though they have the same query volume, which shows that transportation recommendation suffers under-estimate and under-recommend unfairness in certain time periods. Furthermore, as shown in Fig. 1(d), we calculate *recall*, *precision* and *f1-score* for each transport mode in Beijing by the original model [15]. The results show the original model can not give a balanced quality of services for each transport mode and its users, where under-estimate unfairness happened.

3 FASTR Framework

3.1 Overview

The overall workflow of FASTR is shown in Fig. 2, we first input the features mentioned in Sect. 2 to a multi-task wide and deep model for capturing feature co-occurrence relationships. Then, we propose the dual-focal mechanism and the spatiotemporal fairness metrics as well as regularizers to mitigate under-estimate and under-recommend unfairness respectively. Finally, with these well-designed mechanism, metrics, and regularizers, we can have a more balanced quality of recommendation on transport mode for users.

3.2 Multi-task Wide and Deep Learning Model

To capture the co-occurrence and high-order interaction relationships between different features, we adopt the wide and deep learning model [7] that is widely used in many recommender system, and extend it with the multi-task paradigm [16] to serve as our basic model, where the multi-task mechanism improves the performance on minorities and helps to mitigate the under-estimate disparity [9].

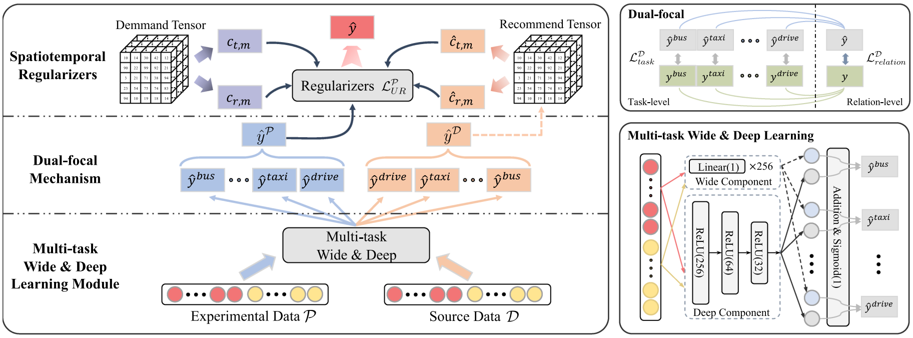


Fig. 2. The overall workflow of FASTR.

Wide and Deep Learning Model. Wide and deep learning consists of a wide component for low-level feature co-occurrence memorization and a deep component for high-level feature co-occurrence generalization. Thus, the wide component with shallow structure is defined as $\hat{y}_i = \mathbf{w}^\top \mathbf{x}_i + b$, where $\mathbf{x}_i$ is the i-th input feature vector, $\mathbf{w}$ is the learnable weighted matrix and b is the bias. The deep component stacks multiple neural network layers to capture higher-order feature representations. Each fully connected layer transform input vector as $\mathbf{z}_{l+1} = \mathrm{ReLU}(\mathbf{w}_l^\top \mathbf{z}_l + b_l)$, where $\mathbf{z}_l$ and $\mathbf{z}_{l+1}$ are the input and output of l-th layer, $\mathbf{w}_l$ and b_l are the weight and bias parameters of layer l. With both wide and deep components, the final prediction of wide and deep learning model can be formulated as $\hat{y}_i = \sigma(\mathbf{w}_w^\top \mathbf{x}_i + \mathbf{w}_d^\top \mathbf{z}_f + b)$, where $\hat{y}_i$ is the final output, σ stands for the activation function, $\mathbf{w}_w$ is the weight parameter of the wide component, and $\mathbf{w}_d$ is the weight parameter of the final output of the deep component $\mathbf{z}_f$.

Multi-task Mechanism. To promote the recommendation performance for users who prefer different transport modes, we follow the settings in [9] who claimed prediction is more accurate when treating the recommendation of each transport mode as an independent task. In particular, we apply the multi-task strategy to divide transportation recommendation into several binary classification tasks that predict whether a user will click on a specific transport mode, where lower-level parameters in the wide and deep components of wide and deep learning model are shared cross all tasks [16]. Notice that we treat the prediction of a transport mode as a binary classification task, where we have 7 tasks (*i.e.*, 7 transport modes) totally in our work. For each transport mode m, the binary classification task of m can be formulated as follows:

$$\hat{y}_i^m = \sigma(\mathbf{w}_w^{m\top}\mathbf{x}_i + \mathbf{w}_d^{m\top}\mathbf{z}_f + b), \tag{1}$$

where $\mathbf{w}_w^m$ and $\mathbf{w}_d^m$ are the task-specific parameters of the wide component and the deep component respectively. And $\hat{y}_i^m \in [0,1]$ indicates the probability of users click on transport mode m.

3.3 Dual-Focal Mechanism for Under-Estimate

As described before, the under-estimate unfairness is usually caused by the performance gap between transport modes in transportation recommendation. Thus, we intuitively need a mechanism that can promote the prediction performance on each transport mode and mitigate the performance gap. In particular, we apply Focal Loss, which has been widely used for computer vision [18] and pays more attention to samples that are more difficult to distinguish for mitigating the under-estimate unfairness. Firstly, we propose task-level focal loss for each binary classification task, which aims to improve the ability of each task on serving the minority users who prefer the task corresponding transport mode but difficult to distinguish. Therefore, we denote task-level focal loss for under-estimate mitigation as the summarization of each binary classification task:

$$\begin{aligned}\mathcal{L}_{task}^{\mathcal{D}} = -\frac{1}{|\mathcal{D}||\mathcal{M}|}\sum_{i\in\mathcal{D}}\sum_{m\in\mathcal{M}} &\big[\alpha_m y_i^m(1-\hat{y}_i^m)^\gamma \log\hat{y}_i^m \\ &+ (1-\alpha_m)(1-y_i^m)(\hat{y}_i^m)^\gamma \log(1-\hat{y}_i^m)\big],\end{aligned} \tag{2}$$

where $\mathcal{M}, \mathcal{D}$ are the set of transport modes and source data respectively. $y_i^m \in \{0,1\}$ is the ground truth that indicates whether user i clicks transport mode m. α_m is the hyperparameter to alleviate binary class imbalance, and γ is the hyperparameter to regulate attentions on the samples that are difficult to distinguish. Taking ground truth y_i^m equals 1 as an example, when the predicted probability $\hat{y}_i^m$ of the i-th sample is nearly 1.0 which means easy to distinguish, the attentions on this sample can be reduced through $(1-\hat{y}_i^m)^\gamma$. And the larger the γ is, the more attention are paid to difficult samples, and vice versa.

Beyond the task-level that focuses on the performance of individual task, we propose relation-level focal loss for mitigating the performance gap between tasks. Specifically, since minority transport modes (*i.e.*, bicycle, taxi and drive) hold less data than the mainstream group, minorities may suffer insufficient training and poor recommendation as described in Sect. 2. Therefore, we apply relation-level focal loss between tasks as follows, where more attention can be paid to minority transport modes.

$$\mathcal{L}_{relation}^{\mathcal{D}} = -\frac{1}{|\mathcal{D}||\mathcal{M}|}\sum_{i\in\mathcal{D}}\sum_{m\in\mathcal{M}} \beta_m y_i^m(1-\hat{y}_i^m)^\gamma \log\hat{y}_i^m, \tag{3}$$

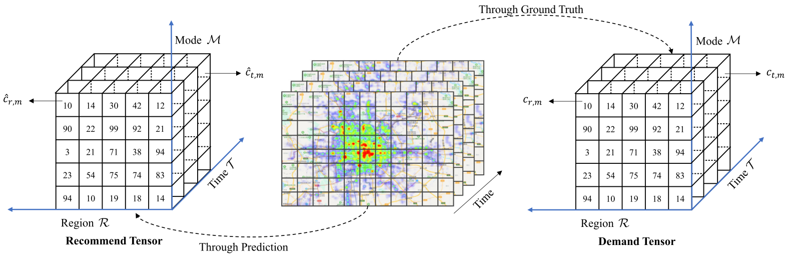


Fig. 3. Demand and recommend tensor construction.

where β_m is the hyperparameter to alleviate multiple class imbalance. With task-level and relation-level focal losses, our dual-focal mechanism can mitigate the under-estimate unfairness by promoting the prediction performance on every transport mode. The overall dual-focal mechanism can be formulated as follows:

$$\mathcal{L}_{UE}^{\mathcal{D}} = \mathcal{L}_{task}^{\mathcal{D}} + \mathcal{L}_{relation}^{\mathcal{D}}. \tag{4}$$

3.4 Spatiotemporal Metrics and Regularizers for Under-Recommend

To mitigate the under-recommend unfairness on recommending lower volume in certain regions and time periods, we first construct demand and recommend tensor in regions and time periods aspects. Then we design a series of spatiotemporal oriented metrics to measure the degree of under-recommend through demand and recommend tensor. And the corresponding regularizers are proposed to mitigate under-recommend unfairness.

Demand and Recommend Tensor Construction. As shown in Fig. 3, we first let $r \in \mathcal{R}$ be the r-th square region of the study area of $\mathcal{R}$, $t \in \mathcal{T}$ be the t-th o'clock in one day, and $m \in \mathcal{M}$ be the m-th transport mode. Then, we calculate the ground truth number of demands to mode m in region r and time t as $c_{r,m}$ and $c_{t,m}$ respectively. Thus we can construct our demand tensor as shown in Fig. 3. What's more, to reflect the under-recommend unfairness of recommender system, we denote $\hat{c}_{r,m}$ and $\hat{c}_{t,m}$ as the recommend volume of recommender system on transport mode m in region r and time t respectively. Then, with the calculated $\hat{c}_{r,m}$ and $\hat{c}_{t,m}$, we formulate recommend tensor in Fig. 3 to proxy recommendation volume.

Region-based Fairness (RF) Metric. Now we formally define our spatial metric RF in the region $\mathcal{R}$ as:

$$RF = P\{(u(r) - u(r')) \leq \epsilon \mid r \neq \ r', \ r, \ r' \in \mathcal{R}\}, \tag{5}$$

where $u(r)$ denotes the degree of under-recommend in region r, and lower $u(r)$ indicates lower under-recommend. And RF can be interpreted as for any two regions, the differences between $u(r)$ and $u(r')$ are not greater than ϵ. To be more direct and remove the interference on selecting ϵ, we modify our RF metric as follows to measure the degree of under-recommend unfairness in the spatial aspect, which is same to Formula 5.

$$RF = \max_{r \in \mathcal{R}}(u(r)) - \min_{r \in \mathcal{R}}(u(r)). \tag{6}$$

Intuitively, we design under-recommend degree of transport mode m in region r as the scaled differences between demands $c_{r,m}$ and recommend volume $\hat{c}_{r,m}$:

$$u(r, m) = \mathrm{ReLU}\left(\frac{c_{r,m} - \hat{c}_{r,m}}{c_{r,m}}\right), \tag{7}$$

where function $\mathrm{ReLU}(\cdot)$ is utilized to filter out the transport mode that not under-recommend. With $u(r, m)$, $u(r)$ can be calculated as follows:

$$u(r) = \frac{\sum_{m \in \mathcal{M}} u(r, m)}{\sum_{m \in \mathcal{M}} \mathrm{sign}(u(r, m))}, \tag{8}$$

where $\mathrm{sign}(\cdot)$ is a function that treats positive numbers as 1 and 0 for others.

Temporal-based Fairness (TF) Metric. Similar to RF, we define temporal metric TF as follows:

$$TF = \max_{t \in \mathcal{T}}(u(t)) - \min_{t \in \mathcal{T}}(u(t)), \tag{9}$$

where TF measures how different the degree of under-recommend over different transport modes from the perspective of time periods. And $u(t)$ can be calculated as follows:

$$\begin{aligned} u(t) &= \frac{\sum_{m \in \mathcal{M}} u(t, m)}{\sum_{m \in \mathcal{M}} \mathrm{sign}(u(t, m))}, \\ u(t, m) &= \mathrm{ReLU}\left(\frac{c_{t,m} - \hat{c}_{t,m}}{c_{t,m}}\right). \end{aligned} \tag{10}$$

Region-Temporal Fairness (RTF) Metric. To measure the whole degree of under-recommend from both spatial and temporal perspective, we formally define RTF, the overall degree of under-recommend, as follows:

$$RTF = \frac{\sum_{r \in \mathcal{R}} u(r)}{|\mathcal{R}|} = \frac{\sum_{t \in \mathcal{T}} u(t)}{|\mathcal{T}|}. \tag{11}$$

Spatiotemporal Regularizer for Under-recommend. Since $\hat{c}_{r,m}$ and $\hat{c}_{r,t}$ can not be calculated directly during model training, we follow the randomized experiments in [1] and collect an experimental data $\mathcal{P}$. Specifically, we rebalance $\mathcal{P}$ to have approximately the same transportation demand distribution at arbitrary regions and time periods. Note that further data restrictions can be applied for alternative goals, and $\mathcal{P}$ is independent of the source data $\mathcal{D}$. Based on RF, TF and experimental data $\mathcal{P}$, we define our spatiotemporal oriented regularizer to constrict under-recommend unfairness:

$$\mathcal{L}_{UR}^{\mathcal{P}} = \lambda_{\mathcal{R}} \frac{\sum_{r \in \mathcal{R}} u(r)}{|\mathcal{R}|} + \lambda_{\mathcal{T}} \frac{\sum_{t \in \mathcal{T}} u(t)}{|\mathcal{T}|}, \tag{12}$$

where $\lambda_{\mathcal{R}}$ and $\lambda_{\mathcal{T}}$ are the weight terms. And the ultimate goal of our mission can be summarized as follows:

$$\mathcal{L} = \mathcal{L}_{UR}^{\mathcal{P}} + \mathcal{L}_{UE}^{\mathcal{D}}. \tag{13}$$

4 Experiments

In this section, we evaluate the performance of our FASTR framework on two real-world datasets described before by the transportation recommendation task.

4.1 Experimental Settings

A) Evaluation Metrics. As described in Sect. 3.4, RF, TF and RTF metrics are utilized to measure the under-recommend degree of our transportation recommender system in the perspective of spatial, temporal and spatiotemporal respectively. Note that lower in RF, TF, and RTF, better in fairness. Besides, we choose to apply macro-recall, variance-recall and maxmin-recall to reveal the performance on mitigating under-estimate unfairness of FASTR in recommending different transport modes. Specifically, the weight of macro-recall for every transport mode is the same, which leads to a fairer evaluation of models. And variance-recall are calculated to measure the differences of performance on predicting different transport modes, where larger the variance is, larger the degree of under-estimate. We also propose the maxmin-recall to describe the difference between the maximum and the minimum recall of transport modes.

B) Baselines & Variants. We compare our approach with four learning-based methods, which are widely used and recognized in the industry, and three variants of our FASTR. Specifically, Logistic Regression (LR) and XGBoost [6] as the most representative models for classification tasks are compared, and the inputs of LR and XGBoost are as same as FASTR. Wide&Deep [7] and DeepFM [11] are two widely acclaimed models for recommendation, who incorporate both shallow and deep relationships between features. Here, we also use the same input as FASTR for them. The ablation study is conducted with three variants defined as follows, 1) FASTR-MR masks dual-focal loss and spatiotemporal regularizers of our FASTR, and we utilize cross-entropy loss for each binary classier, 2) FASTR-MD replaces dual-focal loss with cross-entropy loss, and 3) FASTR-MM masks the multi-task mechanism of our FASTR.

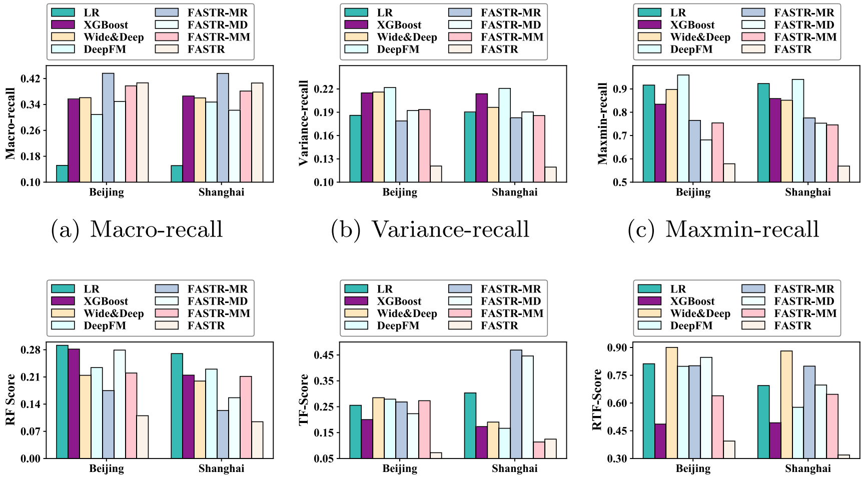
C) Implementation Details. In the implementation phase, we constructed our FASTR by PaddlePaddle[1], which supports a variety of AI-empowered products in Baidu. Specifically, we first transformed categorical features into 32-dimensional embedding vectors, and concatenated them with all other continuous features as the input vector. Then, we fed the input vector into multi-task wide and deep learning model, where the deep component consists of four fully connected layers with 400, 256, 64 and 32 hidden units respectively. And we

[1] https://www.paddlepaddle.org.cn/.

chose to use Sigmoid as our activation function. When implementing our fairness constricts, the class weight α_m and β_m for dual-focal loss were set through balance strategy[2], and the hyperparameter γ was set to 3.0. For $\mathcal{L}_{UR}^{\mathcal{P}}$, we set $\lambda_{\mathcal{R}}$ and $\lambda_{\mathcal{T}}$ both equaled to 0.5. Finally, we set the batch size to 128, learning rate to 5e-4 and trained them with Adam optimizer [21].

4.2 Quantitative Evaluations of FASTR

Performance on Mitigating Under-estimate Unfairness. Figure 4(a), 4(b), and 4(c) show the overall performance on mitigating under-estimate of our FASTR and other methods. And we find three observations through these results. Firstly, as shown in Fig. 4, the macro-recall of our FASTR and its variants are better than other methods, and FASTR achieves much lower variance-recall and maxmin-recall, which means we can provide unbiased transportation recommendation for users without causing too much harm to the overall quality (*i.e.*, macro-recall) compared to other methods. Secondly, FASTR consistently outperforms Wide&Deep and FASTR-MM in terms of macro-recall, variance-recall and maxmin-recall metrics, which proves the effectiveness of multi-task mechanism on mitigating under-estimate unfairness. Thirdly, comparing FASTR and FASTR-MD, the former has better performance on variance-recall and maxmin-recall than the later, which demonstrates the effectiveness of our task-level and relation-level focal loss on mitigating under-estimate unfairness.



(a) Macro-recall (b) Variance-recall (c) Maxmin-recall

(d) RF (e) TF (f) RTF

Fig. 4. Overall performance on transportation recommendation.

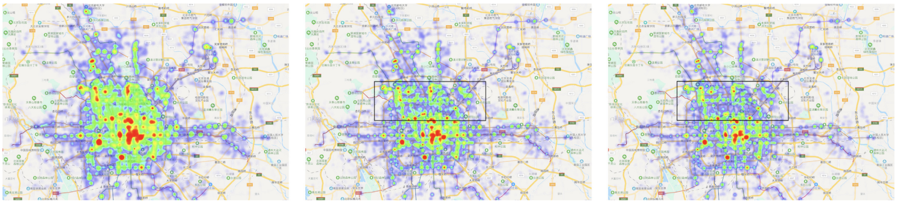
Performance on Mitigating Under-recommend Unfairness. Figure 4(d), 4(e), and 4(f) depict the ability of baselines, variants and our FASTR on mitigat-

[2] https://scikit-learn.org.

ing under-recommend unfairness. And we have three observations through these results. Firstly, our FASTR framework achieves better performance than other methods on RF, TF and RTF metrics, which demonstrates the effectiveness of our spatiotemporal regularizer $\mathcal{L}_{UR}^{\mathcal{P}}$ on mitigating under-recommend unfairness in both region and temporal perspective. Specifically, our FASTR framework beats other methods on RF, TF and RTF up to 18.12%, 21.23% and 50.62% in Beijing, and 17.54%, 17.39% and 56.22% in Shanghai respectively. Secondly, DeepFM and Wide&Deep have similar performance on RF and TF metrics but DeepFM's RTF degree is much higher than Wide&Deep. Since DeepFM has a better fitting ability than Wide&Deep, the bias in datasets may cause this gap in RTF. Comparing FASTR with FASTR-MR and FASTR-MD, we find both spatiotemporal regularizer and dual-focal mechanism are useful to mitigate under-recommend unfairness, where the specially designed spatiotemporal regularizer plays better. Thirdly, we compare FASTR with the best baseline XGBoost and draw Fig. 5 by calculating the distribution of RTF in Beijing. We find that our FARSTR framework recommends more densely than XGBoost, as shown in the black box in Fig. 5(b) and Fig. 5(c), which indicates our FASTR suffers lower under-recommend unfairness.

4.3 The Cost of Fairness in Transportation Recommendation

In this paper, we propose to use dual-focal mechanism with spatial and temporal oriented regularizers to mitigate under-estimate and under-recommend unfairness. However, as shown in Fig. 4, the big improvement in fairness brings performance degradation for the mainstream group. To further reveal the impact of our fairness constraints, we apply our FASTR to an online A/B test in November 2019 in Beijing. And we find there has less than 1% decreasing in overall click-through rate with more than 6% improvement in minorities, which means our FASTR provides fair services for more users. Quantitatively, compared with our original model [15], we have 11.3%, 40.1%, 30.6%, 60.8%, 74.5%, 18.8% improving on macro-recall, variance-recall, maxmin-recall, RF, TF, RTF respectively. The results show that our FASTR is acceptable because of its big improvement in fairness for minority groups but little degradation on performance for the mainstream group, and we can provide fairness-aware transportation recommendation for a better user experience.



(a) Distribution of Queries (b) FASTR on RTF (c) XGBoost on RTF

Fig. 5. Distribution of RTF. (a) queries distribution in Beijing. (b) RTF distribution of FASTR. (c) RTF distribution of XGBoost. Notice that the redder region means lower under-recommend unfairness.

5 Conclusion

In this paper, we investigated the fairness problem in transportation recommendation by mitigating the under-estimate and under-recommend unfairness for users with different travel needs. Specifically, we proposed a Fairness-Aware Spatiotemporal Transportation Recommendation framework (FASTR), which consists of multi-task wide and deep model with dual-focal mechanism for under-estimate unfairness mitigation and tailor-designed spatiotemporal metrics and regularizers for under-recommend unfairness mitigation. Extensive evaluations on real-world datasets validated the effectiveness of our FASTR on mitigating these two types of unfairness, which lead to an unbiased transportation recommendation for users. Besides, through the urban scale A/B test, we confirmed the practicability of our FASTR framework.

Acknowledgement. This research was partially supported by grants from the National Key Research and Development Program of China (Grant No. 2018YFB1402600), and the National Natural Science Foundation of China (Grant No. 91746301, 62072423). And the work was done when the first author interned in Baidu Research.

References

1. Beutel, A., et al.: Fairness in recommendation ranking through pairwise comparisons. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2212–2220 (2019)
2. Beutel, A., Chi, E.H., Cheng, Z., Pham, H., Anderson, J.: Beyond globally optimal: focused learning for improved recommendations. In: Proceedings of the 26th International Conference on World Wide Web, pp. 203–212 (2017)
3. Biega, A.J., Gummadi, K.P., Weikum, G.: Equity of attention: amortizing individual fairness in rankings. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, pp. 405–414 (2018)
4. Calders, T., Verwer, S.: Three Naive Bayes approaches for discrimination-free classification. Data Min. Knowl. Disc. **21**(2), 277–292 (2010). https://doi.org/10.1007/s10618-010-0190-x
5. Chaney, A.J.B., Stewart, B.M., Engelhardt, B.E.: How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In: Proceedings of the 12th ACM Conference on Recommender Systems - RecSys 2018 (2018)
6. Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y.: Xgboost: extreme gradient boosting. R package version 0.4-2, pp. 1–4 (2015)
7. Cheng, H.T., Koc, L., et al.: Wide & deep learning for recommender systems. In: Proceedings of the 1st Workshop on Deep Learning for Recommender systems (2016)
8. Crowson, C.S., Atkinson, E.J., Therneau, T.M.: Assessing calibration of prognostic risk scores. Stat. Methods Med. Res. **25**(4), 1692–1706 (2016)
9. Das, A., Dantcheva, A., Bremond, F.: Mitigating bias in gender, age and ethnicity classification: a multi-task convolution neural network approach. In: Leal-Taixé, L., Roth, S. (eds.) ECCV 2018. LNCS, vol. 11129, pp. 573–585. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11009-3_35

10. Fu, L., Sun, D., Rilett, L.R.: Heuristic shortest path algorithms for transportation applications: state of the art. Comput. Oper. Res. **33**, 3324–3343 (2006)
11. Guo, H., Tang, R., Ye, Y., Li, Z., He, X.: DeepFM: a factorization-machine based neural network for CTR prediction. arXiv preprint arXiv:1703.04247 (2017)
12. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. In: Advances in Neural Information Processing Systems, pp. 3315–3323 (2016)
13. Liu, H., Han, J., Fu, Y., Zhou, J., Lu, X., Xiong, H.: Multi-modal transportation recommendation with unified route representation learning. Proc. VLDB Endow. **14**(3), 342–350 (2021)
14. Liu, H., Tong, Y., Han, J., Zhang, P., Lu, X., Xiong, H.: Incorporating multi-source urban data for personalized and context-aware multi-modal transportation recommendation. IEEE Trans. Knowl. Data Eng. (2020)
15. Liu, H., Tong, Y., Zhang, P., Lu, X., Duan, J., Xiong, H.: Hydra: a personalized and context-aware multi-modal transportation recommendation system. In: Proceedings of the 25th ACM SIGKDD (2019)
16. Ruder, S.: An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098 (2017)
17. Singh, A., Joachims, T.: Fairness of exposure in rankings. In: Proceedings of the 24th ACM SIGKDD (2018)
18. Wang, Z., She, Q., Ward, T.E.: Generative adversarial networks in computer vision: a survey and taxonomy. arXiv preprint arXiv:1906.01529 (2019)
19. Xu, T., Zhu, H., Xiong, H., Zhong, H., Chen, E.: Exploring the social learning of taxi drivers in latent vehicle-to-vehicle networks. IEEE TMC **19**, 1804–1817 (2019)
20. Zheng, Y.: Trajectory data mining: an overview. ACM Trans. Intell. Syst. Technol. (TIST) **6**(3), 1–41 (2015)
21. Zhong, H., et al.: Adam revisited: a weighted past gradients perspective. Front. Comput. Sci. **14**(5), 1–16 (2020). https://doi.org/10.1007/s11704-019-8457-x